Proceedings of IEEE Virtual Reality 2003, pages 259–266, © IEEE Computer Society, 2003.

A Comparative Study of User Performance in a Map-Based Virtual Environment

J. Edward Swan II[*1], Joseph L. Gabbard[2], Deborah Hix[2], Robert S. Schulman[3], Keun Pyo Kim[3]

[1] The Naval Research Laboratory, Washington, DC
[2] Virginia Tech, Systems Research Center, Blacksburg, VA
[3] Virginia Tech, Department of Statistics, Blacksburg, VA

Abstract

We present a comparative study of user performance with tasks involving navigation, visual search, and geometric manipulation, in a map-based battlefield visualization virtual environment (VE). Specifically, our experiment compared user performance of the same task across four different VE platforms: desktop, cave, workbench, and wall. Independent variables were platform type, stereopsis (stereo, mono), movement control mode (rate, position), and frame of reference (egocentric, exocentric). Overall results showed that users performed tasks fastest using the desktop and slowest using the workbench. Other results are detailed below. Notable is that we designed our task in an application context*, with tasking much closer to how users would actually use a real-world battlefield visualization system. This is very uncommon for comparative studies, which are usually designed with abstract tasks to minimize variance. This is, we believe, one of the first and most complex studies to comparatively examine, in an application context, this many key variables affecting VE user interface design.*

Keywords: user-centered design, user interfaces, user interaction, user assessment, usability engineering, usability evaluation, virtual environments, virtual reality, expert heuristic evaluation, formative evaluation, summative evaluation.

1. Background

Collaborative research between the Naval Research Laboratory and Virginia Tech has focused for several years on designing, prototyping, and evaluating user interfaces for virtual environments (VEs). Rather than focusing on developing VEs for technology's sake, we are focusing on developing VEs for their users' sake. We have evolved a sequential usability engineering process (Hix & Gabbard [11]) that is a cost-effective and scientifically-effective progression. The evaluation component of this process involves three phases: performing heuristic evaluation, formative evaluation, and summative evaluation, with iteration as appropriate within and among each. *Heuristic evaluation* is a guidelines-based assessment performed by a user interface design expert. *Formative evaluation* is a user-based assessment with representative users; like a heuristic evaluation, its purpose is to assess and improve a specific user interface. *Summative evaluation,* in contrast, is performed to statistically compare several user interfaces to determine which one is better. This progression leverages the results of each phase by systematically refining the VE user interface. Such evaluations should be a routine part of VE development (Hix and Hartson [13]).

At VR'99 we described our use of heuristic and formative evaluations to study the generic VE user task of navigation, using a battlefield visualization VE called *Dragon* [12]. This paper focuses on our use of the third phase, summative evaluation, with Dragon. Thus, following the VR'99 paper, this is the next chapter in our continuing use and assessment of usability engineering methods, with the dual goal of improving both *product* (here, Dragon) and *process* (here, summative evaluation). This summative study has navigation as a key feature, and also includes the typical VE user tasks of visual search and object manipulation. Notable is that we examined tasks within the real-world *application context* of battlefield visualization. This is different than many human factors studies, which, in order to minimize variance, are set within an abstract context developed specifically to support the evaluation (e.g., Bowman, Johnson, & Hodges [2]; Snow & Williges [18]).

Following related work (section 2), we present our methodology (section 3) and discuss our results (section 4). We conclude with lessons learned (section 5). Our findings should be directly applicable and adaptable by those creating VE applications (e.g., *product results* for VE navigation, visual search, and object manipulation tasks), as well as those undertaking similar VE studies (*process results* for summative evaluations).

We recognize that summative evaluation, as presented here, may appear time-consuming and costly. Such comparative, statistically controlled studies are, indeed, expensive. However, they are critical to validating the science of VE design; without empirically derived findings and guidelines such as those produced by summative studies, VE developers have only their best guesses to design a VE to optimally serve its users' needs. Thus, a key point of our research, and this paper, is to improve and streamline summative evaluations for VEs, so that

[*]Naval Research Laboratory, Code 5580, 4555 Overlook Ave SW, Washington, DC 20375. Email: swan@acm.org

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2003		**00-00-2003 to 00-00-2003**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
A Comparative Study of User Performance in a Map-Based Virtual Environment	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **The Naval Research Laboratory,Washington,DC**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
IEEE Virtual Reality 2003, March 22-26, Los Angeles, California: IEEE Computer Society, 2003

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**8**	

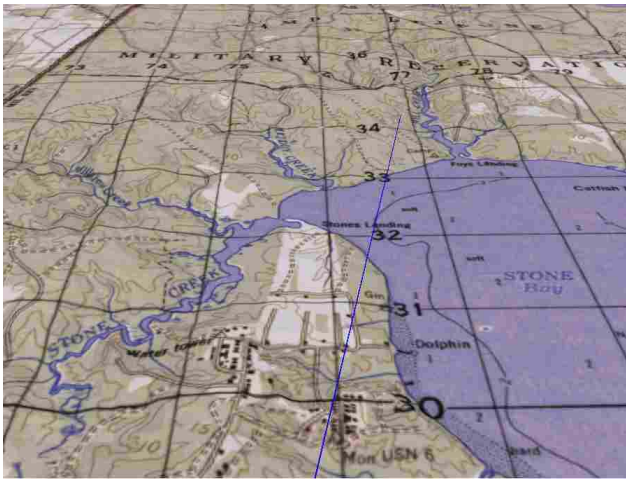

Figure 1: Typical user view of the *Dragon* map during the experiment.

they are more cost-effective and efficient, helping others avoid some of the challenges we encountered.

2. Related Work

Recently, increased interest has emerged in user-centered design and evaluation of VE user interfaces and interaction techniques (Tromp et al. [22; 21]; Johnson [15]; Volbracht & Paelke [23]; Gabbard et al. [8]). We have seen increased interest in usability evaluation for furthering basic VE user interface research, such as providing a scientific means to assess various designs and to test hypotheses (Bowman et al. [2]). We have seen less work presenting usability engineering efforts applied to specific applications and domains, such as Johnson [15], Gabbard et al. [8], Stanney & Reeves [19], and Gabbard, Hix, and Swan [9]. These two different usability engineering contexts (e.g., generic research versus application domain) typically require different types of processes and methods and also provide different types of results (Bowman et al. [1]).

Numerous experiments in the VE community have assessed one or more of the factors we have chosen to study. For example, Hubona et al. [14], Ware & Franck [24], and Zhai et al. [26] examined effects of stereo in VEs on user task performance, especially for positioning tasks. Salzman et al. [17] examined egocentric and exocentric frames of reference to support VE exploration. Card et al. [4] established the widely adopted classification of device design space that includes a discussion of rate- and position-based devices for 2D interaction. Zhai & Milgram [28, 27] examined VE user task performance within this space, researching manipulation tasks using both rate- and position-based interaction techniques. Darken & Sibert [6] examined strategies for wayfinding (the cognitive element of navigation) to study the cues needed for this task in a virtual world. Darken & Cevik [5] also investigated orientation issues and users' frame of reference (ego- and exocentric) with virtual maps during navigation tasks.

We found very few studies that use summative evaluation to empirically examine a large number of experimental factors, which is critical to user task performance in VEs, within an application domain context. Hubona et al. [14] presented a sizable user-based study of depth cues with four factors, one of which is stereopsis. However, this particular study was within a generic context. Goebbels et al. [10] presented a summary of evaluation techniques (including summative) applied to a specific application domain, namely a collaborative medical VE.

3. Method

3.1 Software Application Used in Study

Our study used a three-dimensional map-based virtual environment (VE) derived from the Dragon system [12]. As described in Hix et al. [12], Dragon is a battlefield visualization system that displays a 3D map of the battlespace, as well as military entities (e.g., tanks and ships) represented with semi-realistic models. Dragon allows users to navigate and view the map and symbols, as well as to query and manipulate entities.

As with many VEs, users primarily interact with Dragon using a flightstick: a hand-held, three-button game joystick that we modified by removing its base and placing a six degree-of-freedom tracker sensor inside [7]. The flightstick uses a virtual laser pointer metaphor; a laser beam appears to come out of the flightstick, allowing user interaction with the map or object that the beam intersects. Figure 1 shows a typical user view during our study, with the virtual laser pointer visible as a line extending from the bottom center.

The only user interaction supported by Dragon in our study was the integrated navigation interaction design described in Hix et al. [12, Table 1]. This breaks navigation into three separate modes: (1) *Pan & Zoom* maps the tracker's (x, y, z) degrees of freedom into a corresponding x, y, or z movement of user's eye point, and allows users to pan left & right, up & down, and forward & back; (2) *Rotate* maps the tracker's heading degree of freedom into a rotation of the map around the center-of-interest (COI), where the COI is defined to be the intersection of the virtual laser pointer with the map; and (3) *Tilt* maps the tracker's pitch degree of freedom into rotation around the COI which tilts or pitches the map. Users selected the current navigation mode by pressing one of three buttons on the flightstick.

3.2 Task Performed by Subjects

Subjects performed a series of 17 tasks, each requiring the subject to navigate to a certain location, manipulate the map, and/or answer a specific question based on the map. We called the series of 17 tasks a *task set*, and because platform was a within-subjects variable (see section 3.3),

Task Set A	Task Set C
1. Identify the highway number of the long road running east-west in the upper northeast area of the map.	**1**. Identify the highway number of the long road running east-west in the lower southwest area of the map.
9. Tilt the map so that both the northern horizon and Peru are visible (on the front screen).	**9**. Tilt the map so that both the western horizon and Fulcher Landing are visible (on the front screen).
17. Remaining on the white cube, look around in all directions and indicate which blue object appears farthest from you.	**17**. Remaining on the white cube, look around in all directions and indicate which red object appears farthest from you.

Table 1: Three sample tasks, out of 17 total tasks, shown for two of the four task sets.

we created four task sets (A through D). We designed questions in each task set to be semantically parallel, and therefore functionally equivalent, so that users were performing essentially similar, but *not* identical, tasks on all four platforms. We even made small changes to the map (e.g., adding non-existent towns or bays) to help with this. Table 1 shows the parallel wording of three questions from task sets A and C.

Task set questions fell into thee categories; Table 1 contains an example question from each category. The categories were: (1) *Text tasks*, which involved searching for named items on the map—the subject was either searching for a terrain object to determine its name, or looking for a terrain object when given its name; (2) *Map tasks*, which asked the subject to place the map in a given position; and (3) *Geometric object tasks*, which asked the subject to navigate relative to geometric solids, such as cubes, towers, pyramids, etc.

Subjects began with a training task set, performed on a different map containing similar geographic features. The training task set comprised 7 tasks similar to those in the main task sets.

3.3 Experimental Design

3.3.1 Independent Variables

Our prior work using Dragon had revealed four variables most likely to influence VE tasks such as navigation [12]. Our study manipulated those four independent variables.

Platform (cave, wall, workbench, desktop): For this within-subjects variable, each subject completed a task set using each of four VE display devices: a standard 10' x 10' x 8' cave, a single wall (which we implemented by using only the front screen of the cave), a workbench (with the screen tilted at a 20° angle), and a standard 19" desktop monitor. We made this our only within-subjects variable, because we felt platform results would be the most interesting, and so we wanted the most power for this variable.

Stereopsis (stereo, mono): For this between-subjects variable, half the subjects saw a stereoscopic map and images, while half saw monoscopic. We carefully implemented the stereo and mono conditions to be equivalent — for both conditions we used quad buffering, set the projector configuration to stereo mode, emitted an infrared stereo sync signal, and had all subjects wear stereo liquid-crystal display shutter glasses. This ensured that system performance and users' observed brightness were equivalent for both stereopsis conditions.

Movement control (rate, position): This between-subjects variable describes how a subject's navigational gesture controlled the resulting virtual movement. In *rate* control, the magnitude of the user's gesture controlled the velocity of virtual movement. For example, the user could fly through the virtual world by making and holding a single gesture; the further the user reached, the faster they moved. In *position* control, the magnitude of the user's gesture controlled the distance of virtual movement. For example, to fly through the virtual world the user had to make repeated panning, or ratcheting, gestures, each of which translated the map a distance equivalent to the length of the user's gesture; the further the user reached, the further they moved. Half the subjects used rate control, and half used position control.

Frame of reference (egocentric, exocentric): This between-subjects variable describes whether the user moved themselves through the world (egocentric), or moved the virtual world around themselves (exocentric). With an *egocentric* frame of reference, the user's gesture moved the world in the opposite direction of the gesture, while with an *exocentric* frame of reference, the user's gesture moved the world in the direction of the gesture. Intuitively, in an egocentric frame of reference, the user had the sense of flying an airplane over the map. In an exocentric frame of reference, the user had the sense that the virtual laser pointer was a stick that is used to move the map. Half the subjects saw an egocentric, and half saw an exocentric, frame of reference.

3.3.2 Dependent Variable

Our dependent variable was how long it took subjects to complete each task. To measure this, we set a time stamp when the experimenter began reading a task. During this period, we instructed the subject to listen, and to ask clarifying questions, but not to manipulate the map. When the subject was ready to begin, they gave a verbal response such as "ok" or "now", and we set another time stamp. When the subject completed the task, they gave another verbal response, which was either the name of the item they were searching for, or a phrase such as "ok" or "here it is". At this point, we set another time stamp, which also served as the beginning time stamp for the next task. This simple technique allowed us to factor out the time spent asking questions from the time spent performing tasks.

between subject	Stereopsis		*on*				*off*			
	Movement Control		*rate*		*position*		*rate*		*position*	
	Frame of Reference		*ego*	*exo*	*ego*	*exo*	*ego*	*exo*	*ego*	*exo*
within subject	Platform	*cave*	subjects 1 – 4	subjects 5 – 8	subjects 9 – 12	subjects 13 – 16	subjects 17 – 20	subjects 21 – 24	subjects 25 – 28	subjects 29 – 32
		wall								
		workbench								
		desktop								

Table 2: Experimental design.

3.3.3 Counterbalancing

Table 2 shows our experimental design. We counterbalanced the presentation of the between-subject conditions (stereopsis, movement control, and frame of reference) with a three-way factorial design, yielding eight treatment conditions. We ran 32 subjects, which allowed blocks of four subjects per condition. Each subject performed all four task sets on all four platforms.

With this design, in each 4-subject block, platform presentation order was completely counterbalanced by an order-balanced Latin square. Further, every combination of platform and task set was tested once. For example, within a 4-subject block, task set A was combined with the cave, wall, workbench, and desktop conditions, as were task sets B, C, and D. We chose the training platform to be the same platform subjects used for their first task set, because it was not possible to completely counterbalance training platform within each block. Further, the standard training effect argued that subjects would take the longest to complete the first task set / platform combination, and thus we chose to potentially reduce time from an endpoint of the training effect spectrum rather than from the middle of the spectrum. Our experimental design was non-trivial, and yielded some insights into the properties of 4 x 4 Graeco-Latin squares. The complete design is described in [20].

3.4 Procedure

We had each subject complete a consent form and a questionnaire covering items such as eyesight and experience reading and using maps. Each subject first saw their training platform, where we described map features and taught them how to navigate using the flightstick. After the subject practiced for a few minutes, we described the experimental protocol, and had the subject perform the training task set. After training, we ran the subject through all four platform/task set combinations. The subject next filled out a post-hoc questionnaire, and we then had an informal dialogue about their task strategies, and what they liked and disliked in the user interface.

We always used two experimenters. One led the session, presenting tasks and interacting with the subject, while the other took time stamps. Both experimenters collected additional qualitative data on such items as the subject's navigational strategies, errors, and so forth.

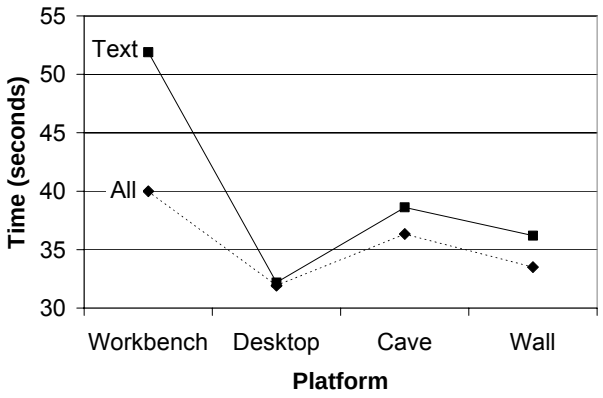

Figure 2: Main effect of platform for all tasks (◆) and text tasks (■).

Subjects did not have problems performing the tasks, nor following the protocol.

3.5 Subjects

Our 32 subjects represented a varied cross-section of civilian (25) and military or retired military (7) personnel. Subjects did not need any special characteristics or skills to participate; all volunteered and received no compensation. We had 6 females and 26 males; ages ranged from 18 to 57 with a median of 29. Our flightstick was designed for right-handed use; 30 subjects reported being right-handed, 1 reported being left-handed and 1 reported being ambidextrous. One subject reported some color blindness. Subjects reported being heavy computer users, averaging more than 20 hours of use per week.

4. Results and Discussion

Our analysis gives results averaged over all 17 tasks, and averaged over the three categories (text, map, and geometric object tasks) discussed in section 3.2.

As shown in Figure 2, there was an *effect of platform* for all tasks ($F(3) = 5.87$, $p < .01$). Subjects performed significantly worse (more slowly) on the workbench than the other platforms. Least squares means (LSM) testing* indicated that the workbench was significantly worse than the desktop ($p < .01$) and the wall ($p < .05$), but not the cave ($p = .300$). There was a similar platform effect for the text tasks ($F(3) = 11.1$, $p < .01$), and LSM testing again indicated the effect came from the poor performance of the workbench compared to desktop ($p < .01$), cave ($p < .01$), and wall ($p < .01$).

While these results appear to be a dramatic condemnation of the workbench, we believe they can mostly be attributed to the projector on our workbench. This projector was visibly fuzzier and dimmer than projectors for our

*All reported least squares means results are tested using Tukey's HSD approach, with appropriate adjustments for within-subject comparisons.

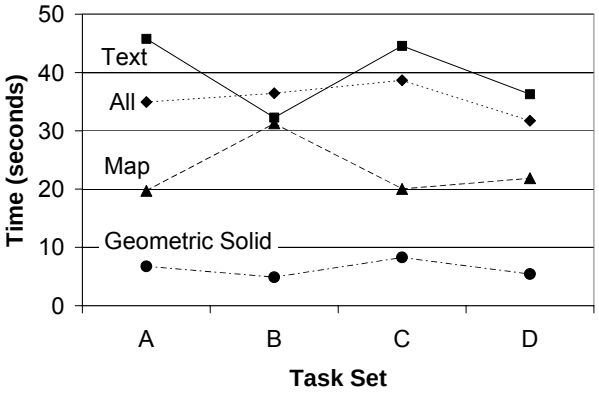

Figure 3: Main effect of task set for all tasks (◆), text tasks (■), map tasks (▲), and geometric object tasks (●).

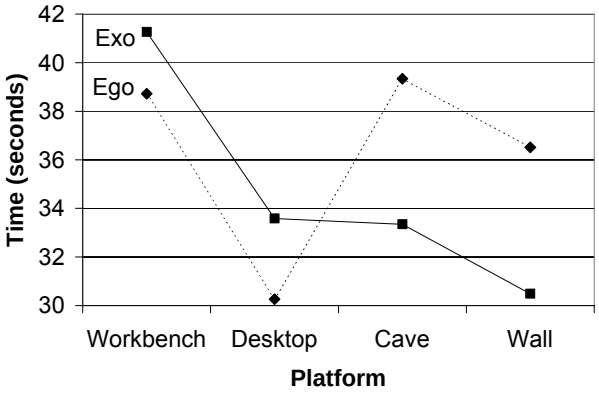

Figure 4: Platform by frame of reference (egocentric (◆), exocentric (■)) interaction for all tasks.

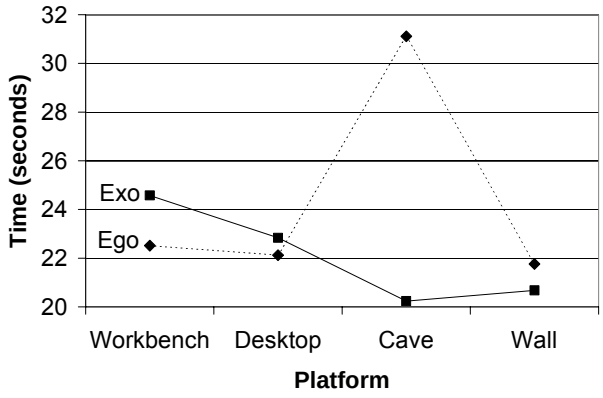

Figure 5: Platform by frame of reference (egocentric (◆), exocentric (■)) interaction for map tasks.

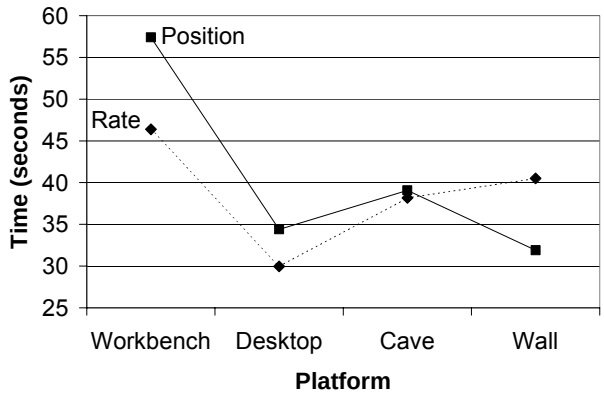

Figure 6: Platform by movement control (rate (◆), position (■)) interaction for text tasks.

cave and wall (which used the same projectors). The desktop display was sharper and clearer than the projectors for any of our VR display devices, largely because it maps the same resolution (1024 x 768 pixels) onto a much smaller surface.

Figure 3 shows the *effect of task set* for all tasks ($F(3) = 3.95$, $p < .05$), text tasks ($F(3) = 6.49$, $p < .01$), map tasks ($F(3) = 9.00$, $p < .01$), and geometric object tasks ($F(3) = 9.09$, $p < .01$). This effect is interesting because we designed the task sets to be as similar as possible, and our intent was that there would be no effect of task set. Table 1 shows that we worded each task in a semantically parallel manner in each task set. However, the tasks turned out not to be parallel in graphical and perceptual areas such as how much distance a user must travel to answer a task, and how easy it is to find a given object. In part, this follows from our decision to run this study within an application rather than a generic context, but the lesson learned is to extensively pilot test scenarios before use.

There was an *effect of task set order of presentation* for all tasks ($F(3) = 27.0$, $p < .01$), text tasks ($F(3) = 16.6$, $p < .01$), map tasks ($F(3) = 11.6$, $p < .01$), and geometric object tasks ($F(3) = 10.6$, $p < .01$). As expected from the standard training effect, subjects became faster as they practiced the task.

Figure 4 shows a *platform by frame of reference interaction* for all tasks ($F(3) = 3.12$, $p < .05$). With the workbench and desktop, subjects performed better with an egocentric frame of reference, while with the cave and wall, they performed better with an exocentric frame of reference. These results are particularly interesting because they refute the common hypotheses that, (1) because the cave encourages a first-person, immersive experience, users would perform better with an egocentric (user moves through the world) frame of reference, and (2) because the workbench encourages a detached, god's-eye overview, users would perform better with an exocentric (user moves the world) frame of reference. These results clearly warrant further study.

A *platform by frame of reference interaction* also occurred for the map tasks ($F(3) = 2.60$, $p = .0601$). Figure 5 suggests, and LSM testing supports, that this interaction is caused by the poor user performance for the egocentric/cave condition. As noted above, this finding refutes

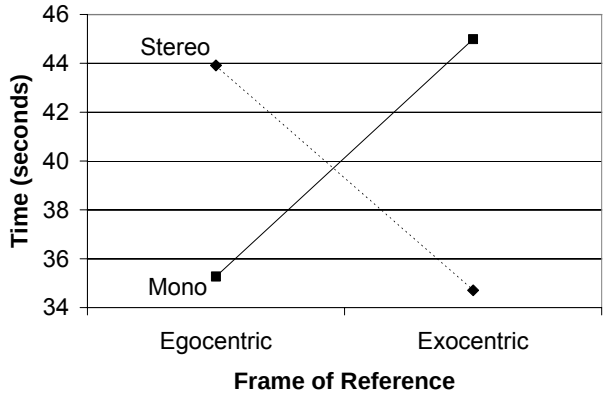

Figure 7: Stereopsis (stereo (◆), mono (■)) by frame of reference interaction for text tasks.

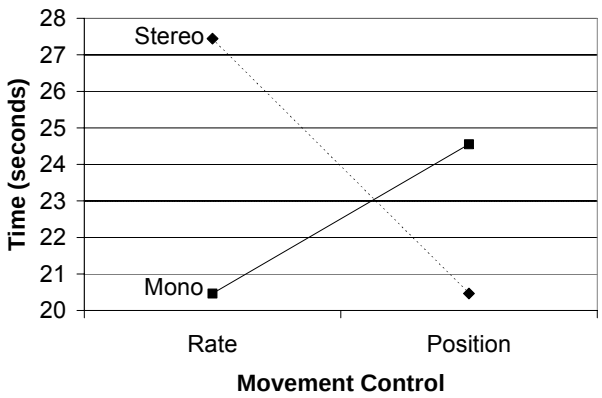

Figure 8: Stereopsis (stereo (◆), mono (■)) by movement control interaction for map tasks.

the expectation that an egocentric frame of reference would yield better performance in a cave.

Figure 6 shows a *platform by movement control interaction* for the text tasks ($F(3) = 2.55$, $p = .0637$). LSM analysis indicates that most of this interaction comes from the poor user performance of the position condition as opposed to the rate condition for the workbench, relative to the position versus rate difference for the desktop, cave, and wall ($p \leq .0324$). We looked, but could not find, examples in the literature which have studied similar factors, so we do not have a basis with which to compare this result.

There was a *stereopsis by frame of reference interaction* for the text tasks ($F(1) = 4.09$, $p = .0545$). Figure 7 shows that mono/egocentric and stereo/exocentric had superior, and similar, performance. Users performed better with an egocentric navigation metaphor when stereopsis was not present. This result is consistent with McCormick et al. [16], where an egocentric frame of reference outperformed an exocentric frame of reference when navigating in a monoscopic scientific dataset.

Stereopsis also had an *interaction with movement control* for the map tasks ($F(1) = 4.11$, $p = .0538$). Figure 8 and LSM testing indicate this interaction is primarily caused by the poor user performance of the stereo/rate condition. Users performed better with position control (as opposed to rate control) when stereopsis was employed. This result is consistent with Zhai [25], where isotonic position control outperformed isometric rate controls in a stereo VR setting (note that in Zhai's framework our flightstick implementation is considered isotonic).

Finally, note that, with the exception of platform, the *lack* of main effects across all tasks is also interesting. There was no effect of stereopsis ($p = .598$), no effect of movement control ($p = .401$), no effect of frame of reference ($p = .682$), nor other interactions. Furthermore there was no effect of platform on map tasks ($p = .366$) and geometric object tasks ($p = .696$). In part, this is a natural result of significant interactions, and in part it reflects the conservatism of the Tukey approach. But it also highlights the subtlety of such comparisons, even with careful attention to experimental design, collaboration with experienced statisticians, and a relatively large number (32) of subjects. We expect the main effect of stereopsis may not have been strong because we evaluated user task performance in a 3D terrain that is essentially a flat map; the Dragon application is targeted for coastal terrain (see Figure 1). We initially anticipated main effects and interesting interactions, in particular with platform type, from the other variables as well.

Our data have adequate power to generate significant main effects and interactions even for single tasks [20], and some of these results are intriguing. However, giving complete results is beyond the scope of this paper; we will report these in future publications.

5. Lessons Learned and Future Work

From this study we have attempted to abstract high-level findings of interest to the VE and usability communities. Here we single out two issues important to continued work in usability engineering, especially of VEs, as well as some other issues, all warranting further investigation.

5.1 Distilling Display Requirements

A striking aspect of our results is the poor performance of the workbench; especially for text tasks. Our reasoning is that this was caused by the fact that the workbench display was noticeably fuzzier and dimmer than the other displays. Although we knew this when preparing for the experiment, we purposefully chose to proceed with the study to see how legacy, aging, inferior, or slightly dated equipment would fare against more current displays.

Notice that the desktop had the best time of all platforms. Many of our tasks required finding, identifying, and/or reading text or objects labeled with text. While all displays were set to 1024 x 768 pixels, the size of the projection surface varied enough to conjecture that pixel density is more critical than field of view or display size. Our

observations and qualitative data support this claim. For example, evaluators often noted that desktop users did not have to zoom in to complete tasks involving reading text.

Since the 1960s, many characteristics of computers have increased or decreased by orders of magnitude, such as speed, memory size, cost, and weight. Interestingly, display density has not. This problem is confounded in VEs because large displays use the same number of pixels as monitors, but spread the pixels over larger areas to support immersion. This research suggests we should further research user task performance using high resolution displays.

What is also interesting is that, for many of our tasks, statistically analyzed across a number of interesting parameters and task types, we found no differences. Furthermore, we found *no effect of platform* at all in map tasks and geometric object tasks. This begs examination of the important question: "Why are we building large-display VEs and incurring the resulting expense if the user benefit is not there?"

5.2 Designing User Tasks to Support Comparative Studies in Application Contexts

Designing usability evaluation studies within an application context presents a number of challenges that are not present in generic, basic-research user studies [1]. As we considered our statistical results, we noted considerable variance that may be hiding otherwise notable findings. We suggest that much of this variance comes from uncontrollable unknowns associated with the increased complexity of application software, relative to abstract testbed software that is developed only for research and experimentation.

Further, the types of user tasks needed to support comparative studies within application contexts are very different than those appropriate in generic user studies. We suggest that designing user tasks for an application-based study force a tradeoff between the need to obtain qualitative usability results as well as the need to truly represent end-user tasking *versus* the need to obtain clean, powerful, extensive statistical findings.

For example, abstract testbed user studies typically employ atomic tasks that have clear starting points and ending points, and which are designed specifically to reduce variance and thereby obtain precise statistical results. In application settings, atomic tasks are strung together to establish higher-level real-world tasking, and thus create a high number of task dependencies. In creating scenarios representative of real-world user tasking, it is difficult to neatly divide and constrain singleton tasks, at least in a meaningful fashion. The result is user tasks that may be, for example, hard to precisely time. Results can then vary due to user strategy and previous system state, rather than because of specific independent variables manipulated in the study.

This tradeoff can be best managed by identifying task sequences and dependencies in advance so that the type of data collected and strategies for collecting data can be designed accordingly. For example, design tasks and types of task responses to maximize quantitative data collected within the constraints of typical application-specific tasks sets. It is also useful to design task sets that have clear closure (end of task) for both evaluator and user, but without compromising the representative application-level task flow. This can be done by designing and timing small task sets rather than singleton tasks. Another strategy is to design tasks where users have to achieve a certain level of competence before proceeding. While this may help ensure a constant starting point for subsequent tasks, timing strategies will need to take into account the additional time needed to establish the level of competence.

Our application-specific approach to usability evaluation may indicate why we do not have clean, expected, widespread statistical results — the complexity of the application and user tasks introduced variance. Our user tasks were necessarily higher-level tasks, perhaps more appropriate for a qualitative analysis. The alternative was to use focused, atomic tasks that might lead to strong statistical results, but those results may not be widely applicable to real-world application domains.

5.3 Future Work

We expect to continue exploring those VE characteristics which significantly effect task performance within VE applications (e.g., user task performance using high resolution displays), as we also continue evolving usability engineering approaches to support such exploration.

As mentioned in section 4, some of our single-task analysis results suggest that there may be some interesting interactions between stereopsis and movement control, as well as stereopsis and frame of reference. These results could be further explored by performing a focused study on these variables within an application context.

Our results further suggest that different navigation paradigms may be needed for different navigation situations. For example, our current technique was optimized for large-scale navigation over a terrain map [12]. Other combinations of factors may be more appropriate for finer-scaled tasks such as streetwise navigation in a city, navigating abstract spaces such as scientific data, manipulating near-field objects, and so forth.

Acknowledgements

Dr. Mike McGee and Eric Nash helped in the early stages of experimental design for this experiment. We are most grateful to Greg Schmidt and Doug Maxwell, who helped run subjects. Also at NRL, Dr. Larry Rosenblum gave resources and guidance. This research was funded by the Office of Naval Research under Program Managers Dr. Helen Gigley, Dr. Astrid Schmidt-Nielsen, and Dr. Paul Quinn. We would like to thank Dr. Gigley and Dr. Schmidt-Nielsen for their continued support of an on-

going synergistic collaboration in human-computer interaction research between Virginia Tech and NRL over the past decade.

References

[1] Bowman, D, Gabbard, JL, Hix, D, (2002) "A Survey of Usability Evaluation in Virtual Environments: Classification and Comparison of Methods". *Presence: Teleoperators and Virtual Environments*, 11(4), pp. 435–455.

[2] Bowman, D, Johnson, D., and Hodges, L. (1999). "Testbed Evaluation of VE Interaction Techniques". Proc. *ACM Symposium on Virtual Reality Software and Technology*, pp. 26–33.

[3] Bowman, D, Koller, D, and Hodges, LF. (1998). "A Methodology for the Evaluation of Travel Techniques for Immersive Virtual Environments". *Virtual Reality: Journal of the Virtual Reality Society*, 3, pp. 120–131.

[4] Card, SA, Mackinlay, JD, and Robertson, GG (1991). "A Morphological Analysis of the Design Space of Input Devices". *ACM Transactions on Information Systems*, 9(2), pp. 99–122.

[5] Darken, RP and Cevik, H. (1999). "Map Usage in Virtual Environments: Orientation Issues". Proc. *IEEE Virtual Reality '99*, pp. 133–140.

[6] Darken, RP and Sibert, JL. (1996). "Wayfinding Strategies and Behaviors in Large Virtual Environments". Proc. *Human Factors in Computing Systems (CHI '96)*, pp. 142–149.

[7] Durbin, J, Swan II, JE, Colbert, B, Crowe, J, King, R, King, T, Scannell, C, Wartell, Z, and Welsh, T. (1998). "Battlefield Visualization on the Responsive Workbench". Proc. *IEEE Visualization '98*, pp. 463–466.

[8] Gabbard, JL, Swan II, JE, Hix, D, Lanzagorta, M, Livingston, M, Brown, D, and Julier, S, (2002). "Usability Engineering: Domain Analysis Activities for Augmented Reality Systems". *The Engineering Reality of Virtual Reality 2002, Proceedings of SPIE Vol. 4660*.

[9] Gabbard, JL, Hix, D, and Swan II, JE. (1999). "User Centered Design and Evaluation of Virtual Environments", *IEEE Computer Graphics and Applications*, 19(6), pp. 51–59.

[10] Goebbels, G, Augustin, S, Lalioti, V, (2001). "Co-Presence and Co-Working in Distributed Collaborative Virtual Environments", Proc. *1st Conference on Computer Graphics, Virtual Reality and Visualization*, pp. 109–114.

[11] Hix, D, Gabbard, JL, (2002). "Usability Engineering of Virtual Environments". In Stanney, K. (Ed.), *Handbook of Virtual Environments: Design, Implementation and Applications*, Lawrence Erlbaum Associates, pp. 681–699.

[12] Hix, D, Swan II, JE, Gabbard, JL, McGee, M, Durbin, J, King, T, (1999). "User-Centered Design and Evaluation of a Real-Time Battlefield Visualization Virtual Environment". Proc. *IEEE Virtual Reality '99*, pp. 96–103.

[13] Hix, D. and Hartson, HR, (1993). *Developing User Interfaces: Ensuring Usability Through Product & Process*. John Wiley & Sons, Inc.

[14] Hubona, GS, Wheeler, PN, Shirah, GW, and Brandt, M, (1999), "The Relative Contributions of Stereo, Lighting, and Background Scenes in Promoting 3D Depth Visualization". *ACM Transactions on Computer-Human Interaction (TOCHI)*, 6(3), pp. 214–242.

[15] Johnson, C (1999). "Evaluating the Contribution of Desktop VR for Safety-Critical Applications". Proc. *SAFECOMP 1999*, pp. 67–78.

[16] McCormick, E, Wickens, CD, Banks, R, and Yeh, M, (1998), "Frame of Reference Effects on Scientific Visualization Subtasks". *Human Factors*, 40(3), pp. 443–451.

[17] Salzman, MC, Dede, C, and Loftin, RB, (1999). "VR's Frames of Reference: a Visualization Technique for Mastering Abstract Multidimensional Information". *ACM Special Interest Group on Computer-Human Interaction (SIGCHI '99)*, pp. 489–495.

[18] Snow, MP and Williges, RC, (1998). "Empirical Models Based on Free-Modulus Magnitude Estimation of Perceived Presence in Virtual Environments". *Human Factors*, 40(3), pp. 386–402.

[19] Stanney, KM and Reeves, L. (2000). "COVE Evaluation Report". Final Report, Contract No.N61339-99-C-0098, Orlando, FL: Naval Air Warfare Center – Training Systems Division, 7/00.

[20] Swan II, JE, Gabbard, JL, and Hix, D, (2002). "Dragon Navigation: A Study of User Performance in a Virtual Environment with Four Independent Variables". *Naval Research Laboratory Memorandum Report* (in preparation).

[21] Tromp, J, Steed, A, Kaur, K (1999). "Systematic Usability Design for Virtual Environments". *ACM Symposium on Virtual Reality Software and Technology (VRST '99),* pp. 20–22.

[22] Tromp, J, Hand, C, Kaur, K, Istance, H, and Steed, A. (1998). "Methods, Results and Future Directions", Proc. *First International Workshop on Usability Evaluation for Virtual Environments*; 17 December 1998, De Montfort University, Leicester, UK.

[23] Volbracht, S, Paelke, V. (2000). Workshop on *Guiding Users through Interactive Experiences: Usability Centered Design and Evaluation of Virtual 3D Environments*, Paderborn, Germany, April 13-14, 2000.

[24] Ware, C and Franck, G, (1996). "Evaluating Stereo and Motion Cues for Visualizing Information Nets in Three Dimensions". *ACM Transactions on Graphics*, 15(2), April 1996, pp. 121–140.

[25] Zhai, S, (1998). "User Performance in Relation to 3D Input Device Design". Proc. *ACM SIGGRAPH*, 32(4), November 1998, pp. 50–54.

[26] Zhai, S, Buxton, W, and Milgram, P, (1994). "The Silk Cursor: Investigating Transparency for 3D Target Acquisition". Proc. *Human Factors in Computing Systems (CHI '94)*, pp. 459–464.

[27] Zhai, S. and Milgram, P., (1994). "Input techniques for HCI in 3D Environments". Proc. *Human Factors in Computing Systems (CHI '94) Conference Proc. Companion*, pp. 85–86.

[28] Zhai, S and Milgram, P, (1993). "Human Performance Evaluation of Manipulation Schemes in Virtual Environments". Proc. *IEEE Virtual Reality Annual International Symposium (VRAIS '93)*, pp. 155–161.